Variational Methods for Latent Variable Problems

Ryan Giordano (for Johns Hopkins Biostats BLAST working group)

Oct, 2021

Massachusetts Institute of Technology

Outline

Outline for today:

- Some examples of latent variable models
- A template: The Neyman-Scott "paradox" and marginalization
- Bayesian versus frequentist approaches to marginalization
- The classical EM algorithm (in brief)

Next week, we will build on these ideas to present more general variational inference.

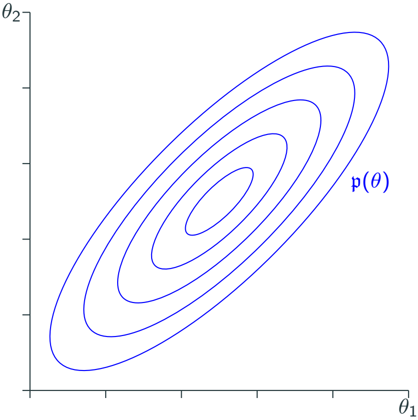
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



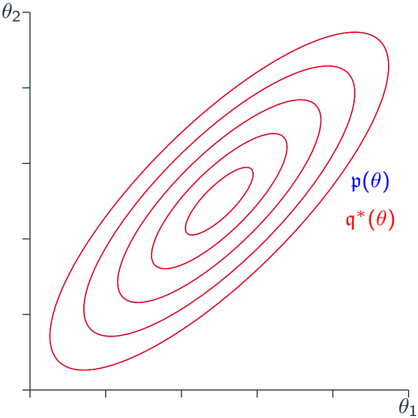
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) =$$
$$- \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficiently expressive families recover the target distribution.

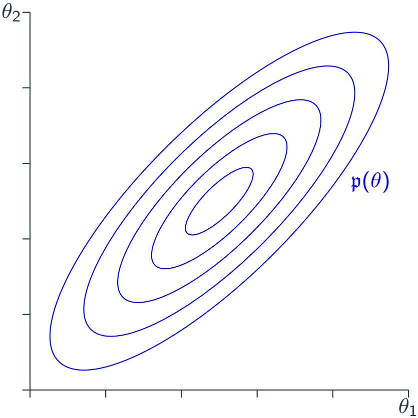
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?

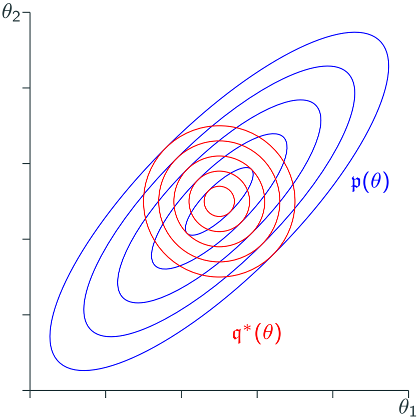


$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\,\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



KL minimizers "fit inside" the second argument.

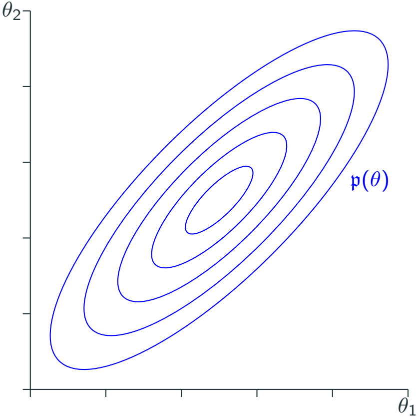
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) = \\ -\underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right] + \underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?

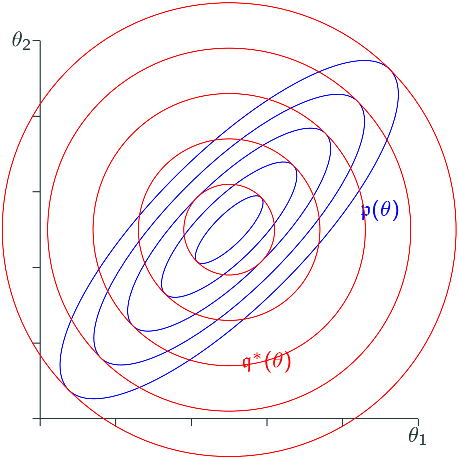


$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}\left[\log\mathfrak{q}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}\left[\log\mathfrak{p}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?



KL minimizers "fit inside" the second argument.

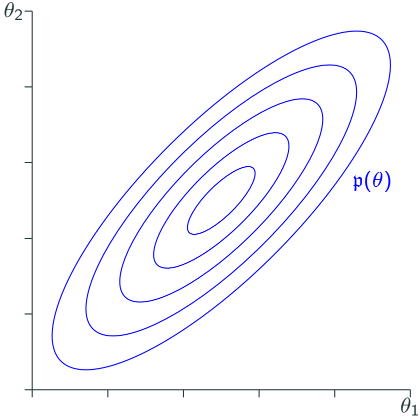
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) =$$
$$- \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$
$\underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \left(- \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]} \right)$?



KL divergence exercises

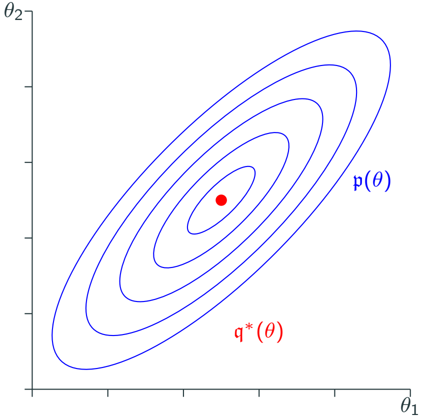
$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$

$$\underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \left(- \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]} \right)?$$



Without the entropy, the KL minimizer concentrates on the maximum of $\log \mathfrak{p}(\theta)$.

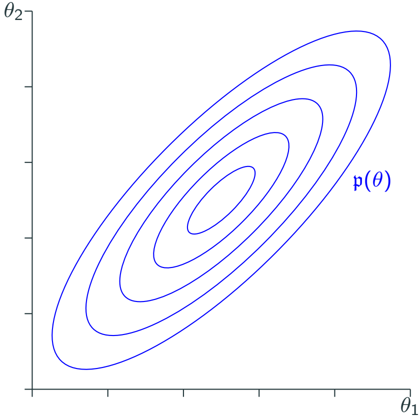
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{Bivariate normals}\}$

What is $\mathfrak{q}^*(\theta) =$
$$\operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}}\left(-\cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right]} + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]\right)?$$



KL divergence exercises

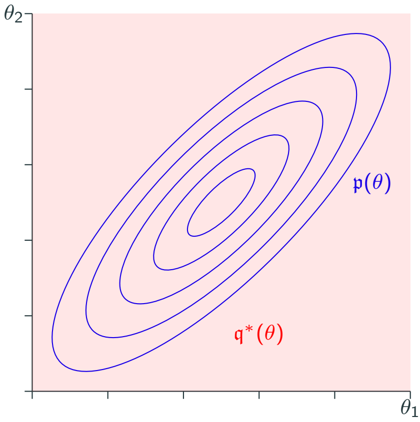
$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$

$$\operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \left(- \cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right]} + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right] \right)?$$



Without $\log \mathfrak{p}(\theta)$, the KL minimizer is infinitely dispersed.

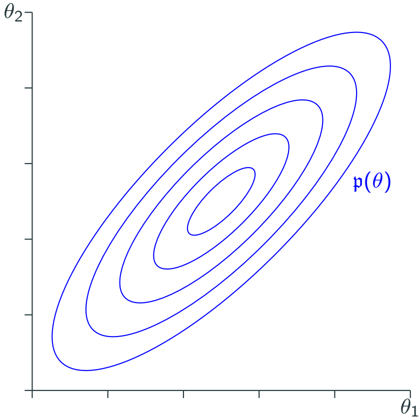
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Point masses}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



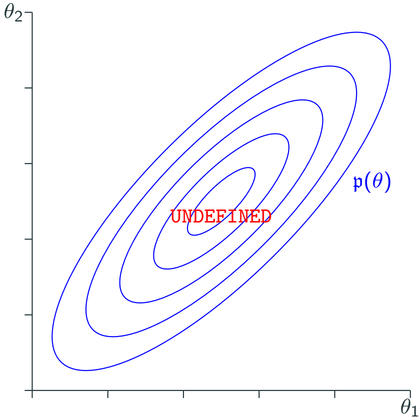
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}[\log \mathfrak{p}(\theta)] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}[\log \mathfrak{q}(\theta)]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Point masses}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Without a common dominating measure, the KL divergence is undefined.

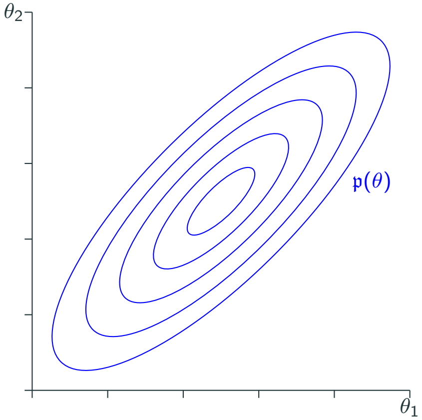
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



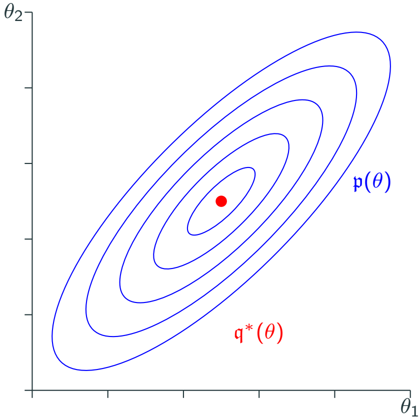
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficently concentrated distributions with constant entropy act like a point mass at the maximum of $\log \mathfrak{p}(\theta)$.

Conclusions